Test Guideline (TG) Page (pp) and Paragraph (¶)	Comments
General	The data generated in the OECD validation program demonstrated the ability of the rat uterotrophic bioassay to reproducibly detect a small number of estrogenic substances when laboratories were instructed to use specific doses for each non-coded test article in one of four assay protocols. Because OECD test guidelines should be based on adequately validated test methods (OECD GD 34), the validation database would appear to be insufficient for this purpose. Therefore, it is recommended that the Uterotrophic Bioassay be issued as an OECD Guidance Document that could be used as the basis for further studies that could lead to an adequate demonstration of validation; such tests could be conducted as part of a U.S. EPA or OECD testing program and could use the ICCVAM recommended list of reference substances for estrogen and androgen receptor activity as a basis for comparative performance. It is also recommended that a comprehensive retrospective evaluation be conducted that integrates the OECD validation study data with historical data to better define the performance characteristics and the limitations of the test method, and to identify (any) data gaps that would need to be filled In any test guideline developed for the uterotrophic bioassay, it is strongly recommended that the context for utilization of the data from this test method within a regulatory framework be clearly stated. For example "the uterotrophic bioassay is intended to be used only as a screen to identify substances with potential estrogenic activity (i.e., estrogen agonists) as part of a battery of in vitro and in vivo tests to identify substances with the potential to interact with the endocrine system."
Title	The word "(anti)" and inclusion of this term throughout the main body of the text should be deleted, as the ability of the uterotrophic assay to detect anti-estrogens has not been characterized. Also, the validation study utilized rats only and unless data are cited and an independent review of that data conducted that concludes that rats and mice can be used interchangeably, the protocol should be restricted to rats.

Test Guideline (TG) Page (pp) and Paragraph (¶)	Comments
pp. 1, ¶ 1	This paragraph states that "extensive" intra- and interlaboratory studies were conducted. However, considering the number of chemicals tested, the word "extensive" is inappropriate or needs to be defined. The word "both" in line 6 should be deleted.
pp. 1, ¶ 2	This paragraph states that the Uterotrophic Bioassay was first standardized and validated by an expert committee in 1962 (TG reference no. 33, "R.I. Dorfman. Standard Methods Adopted by Official Organizations. New York, Academic Press (1962)). This reference is incomplete, our search indicates that it most likely refers to a section in a series of monographs published in 1962 and edited by R.I. Dorfman entitled "Methods in Hormone Research, Vol. II, Part IV: Standard Methods Adopted by Official Organizations." If this is the correct reference, it only specifies the protocols to be used in mouse uterotrophic bioassays, with doses administered by oral gavage or subcutaneous injection. There are no data supporting validation of either test method in Part IV. Indeed, the protocols specify the use of 4 dose groups with 12 animals per dose group, in contrast to the draft TG, which specifies the use of 2 dose groups with 6 animals per group. Considering these facts, the inaccurate referral to an earlier standardization and validation effort by an expert committee in 1962 should be eliminated.
pp. 1, ¶ 4	This paragraph states that no false negatives should be allowed for screening tests. However, due to the few chemicals tested in the validation study, the false negative rate for this assay using the validation study protocol is undefined. Using fewer dose groups and fewer animals per dose group than the standardized protocol for screening in Dorfman (1962) would be expected to result in fewer positives.
pp. 1, ¶ 5	It is difficult to see how identifying meaningful "negatives" posed a problem as there are many substances without estrogenic activity. The phrase "due to … for this purpose" should be deleted.

Test Guideline (TG) Page (pp) and Paragraph (¶)	Comments
pp. 2, ¶ 7	This paragraph states that "This Guideline is based on those protocols employed in the OECD validation study which have been shown to be reliable and repeatable in intra- and interlaboratory studies (5)(7)." First, the structure of the sentence implies that there were protocols which were not reliable or repeatable. Second, in Table 2 of the US submission to the OECD in regard to the first version of the proposed TG, the US position states "Also, agree that the ability of laboratories to test coded substances, to select appropriate doses, and to obtain reproducible and accurate results using the complete test method protocol has not been demonstrated." This statement by the US National Coordinator clearly indicates the shortcomings of this particular validation study and brings to question the validity of the statement in this paragraph. This paragraph further states that it was shown in the OECD validation study that the ovariectomized (OVX) adult female rat and the immature non-OVX female rat methods have equal sensitivity. While an accurate statement (for the validation study), the database from this validation study is too limited to support this as a general blanket statement in terms of saying these two methods are equal. This paragraph states also that the immature rat has an intact HPG axis, making the test system less specific but covering a larger scope of investigation because (in contrast to the OVX female method), it can respond to substances that interact with the HPG axis rather than just the ER. This additional information indicates that either the immature system should not be used because of the increased likelihood of obtaining a false ER response or should be used because the test method also detects substances that interact with the HPG axis, which is also important information. Regardless, this statement means, on face value, that the two model systems differ in sensitivity. The TG should recommend one method and provide a justification for that recommendation, or more clearly state the advantages and limitations of the two methods, as a screening assay for detecting substances with estrogenic activity. The paragraph states that the uterotrophic response is not entirely of estrogenic origin, and that certain non-estrogenic steroids and synthetic progestins may also lead to a stimulative response. The TG states that "Any response may be analyzed histologically for keratinization and cornification of the vagina." Considering the purpose of this test is to identify substances with estrogenic activity, it would seem that this step is necessary to make the test method more accurate. Also, this paragraph states that any positive outcome should normally initiate actions for further clarification by the use of other *in vivo* or *in vitro* assays. Why would that occur if, as Paragraph 4 states, this test method "is embedded within a battery of tests…". Furthermore, if in vitro and different

Test Guideline (TG) Page (pp) and Paragraph (¶)	Comments
pp. 3,¶ 10	The TG acknowledges that the Uterotrophic Bioassay is not validated for the screening of antiestrogenic substances and therefore data, especially negative, derived from "this procedure of the assay" should be interpreted cautiously. Since the TG is now only for the screening of agonist substances, the mention of a "cautious interpretation" of negative data should be deleted from this paragraph.
pp. 3, ¶ 14	This paragraph states that the "Uterotrophic Bioassay relies for its sensitivity on an animal test system in which the hypothalamic-pituitary-ovarian axis is not functional." This statement is inconsistent with the statement in paragraph 7 that immature animals have an active HPG axis. See also paragraph 31.
pp. 4, ¶ 17	The TG states that Sprague-Dawley and Wistar rat strains were used during validation but also states that other commonly used laboratory rodent strains may be used. The TG should recommend that these two strains should be used and that the use of other strains must be adequately justified (e.g., by providing comparative sensitivity data). This approach will minimize an unnecessary variable in the test method and will help to clarify what constitutes an acceptable phytoestrogen level in the diet (paragraph 23).
pp. 4, ¶ 18 and 19	The last sentence in paragraph 18 should state "Healthy animals should be employed." since paragraph 17 deals with strains. This paragraph states that it is "biologically plausible" that rats and mice will have virtually identical responses in the adult OVX and immature Uterotrophic Bioassay. "Biological plausibility" is not sufficient enough to support the use of mice in the TG. Believing in the use of this species for this assay does not constitute a scientifically supportable position and as such the mouse should not be endorsed without bridging data and an independent evaluation of that bridging data that clearly indicates sufficient similarity of response between the two species. If mice are included, then any specific protocol changes must be provided and scientifically justified.
pp. 5, ¶ 22-25	The information provided in these sections is likely to result in the unnecessary use of animals (i.e., waiting until unexpected results are obtained to conduct an analysis). From a humane use of animals perspective, it is important that only diets and bedding demonstrated (either analytically or biologically) to be suitable for this test be used.

Test Guideline (TG) Page (pp) and Paragraph (¶)	Comments
pp. 5, ¶ 23	ICCVAM questions the use of 350 microgram genistein equivalents/gram as the limit for laboratory diet. This is contradicted by current literature (Thigpen et al. 2004), which specifically states that diets containing less than 325 to 350 micrograms/g TGE still have the potential to alter the results of vaginal opening and uterotrophic assays. This continues to be an unresolved issue, and is especially critical to the detection of weak acting estrogenic substances. Specifying the strains used as being Sprague-Dawley or Wistar would eliminate this concern.
pp. 6, ¶ 29	This paragraph is confusing as paragraph 17 states that "The laboratory should demonstrate the sensitivity of the strain used, e.g. by including appropriate positive control groups in its assay." Paragraph 17 seems to state that, regardless of the strain used, a concurrent positive control should be included in each test. While there is appreciation that the use of a concurrent positive control increases the numbers of animals per study, given the potential variations in this test method associated with e.g. diet, bedding, the presence of estrogenic tissue in OVX females, the onset of puberty among some immature females, and the fact that a failed quality control test conducted periodically would necessitate all studies conducted after the last qualifying test be discarded, it appears critical that a concurrent weak-acting positive control (or a weak acting dose of a potent positive control) be included in each study. Once sufficient historical control data had been collected, a laboratory could demonstrate that a concurrent positive control is not needed. See also section 42
pp. 7, ¶ 34	The use of post-operation analgesics should be considered; this is required to reduce pain and suffering associated with the trauma of the operation unless it can be shown to adversely affect the assay.
pp. 8, ¶ 36	This paragraph deals with the selection of dose groups. Stating that a dose level that induces a significant uterotrophic effect is one of the selection criteria for the highest dose level is inappropriate. Making this a separate statement in terms of experimental outcome would be appropriate (e.g., a study conducted at the MTD, the limit dose, or at a dose level that induced a positive uterotrophic response would be accepted).

Test Guideline (TG) Page (pp) and Paragraph (¶)	Comments
pp. 8, ¶ 38	This paragraph states that the maximum limit dose should be 1000 mg/kg/dy. However, other OECD short-term in vivo test guidelines (specifically those for genetic toxicology) mandate 2000 mg/kg/dy for studies of less than 14 days. It is not clear why the dose level for these studies are different, especially based on the likelihood that infants and children may be especially sensitive to endocrine disruption. Furthermore, as different regulatory agencies have different limit dose requirements, the statement "The limit test applies except when human exposure data indicates the need for a higher dose level to be used." should be revised to state "The limit test applies except when there is a specific regulatory mandate that a higher dose level be tested, or when human exposure data indicates the need for a higher dose level to be used."
pp. 8, ¶ 39	This section states that range finding results can be used to select an acceptable maximum and lower doses and recommend the number of dose groups. As only two dose groups are stated to be needed, should this paragraph not say "select the acceptable maximum and minimum dose groups"?
pp. 8, ¶ 40	This section should instruct experimenters to provide a rationale to justify the route of administration.
pp. 9, ¶ 41	This section states that dosing of OVX rats may extend up to seven days, but in the validation experiments there was no significant or consistent advantage over the three-day treatment. Therefore, to simplify the protocol, the TG should recommend that animals be dosed for three days only.
pp. 9, ¶ 42	The TG should recommend as a concurrent positive control a substance (or a dose level of a strong agonist) that induces a relatively weak but consistently statistically positive response. Failure to detect a positive but weak response would necessitate excluding the study (or all studies since the last qualifying test) (see comments for pp. 5, ¶ 22-25 above).
pp. 9, ¶ 46	Unless justified, the optional weighing of feeders to measure food consumption should be deleted. It should be noted that immature animals are group housed which affects the reliability of this measurement.
pp. 10, ¶ 52	The purpose for the optional investigations (histopathology on the uterus and/or vagina) should be provided, as well as how such data are to be used.
pp. 11, ¶ 53 - 55	Data collected and reported should include data generated from the procedures used to establish and maintain performance standards (historical database) and used to control for possible phytoestrogen content in feeds and bedding (see comments for pp. 5, ¶ 22-25 above).
pp. 11, ¶ 54	The TG should state whether statistical tests are one-tailed or two-tailed in testing for a significant increase in uterine weight.

Test Guideline (TG) Page (pp) and Paragraph (¶)	Comments
pp. 12, ¶ 55	Under test animals, stating source in the first bullet and supplier in the second bullet is redundant unless source refers to where the strain was first derived.
pp. 13, ¶ 56	Statistical significance should not be the sole determinant for the toxicological relevance of an observation and that performance standards should be established using an X fold induction (as measured by dividing the maximal response yielded from the positive control by the response from the negative control).
pp. 13, ¶ 59	Acceptance criteria for control uterine weights should be defined when establishing performance standards with the historical database (again, see comments for pp. 5, ¶ 22-25 above).
pp. 13, ¶ 60	As stated earlier, concurrent positive controls should be included.
pp. 14, ¶ 62	Based on the statement that blotted uterine weights show less variability than wet uterine weights, and that blotted weights are to be given preference for the final interpretation, a justification for measuring wet uterine weights should be provided or the protocol should just state that blotted weights are to be used.
pp. 14, ¶ 63	See comments on pp. 2,¶ 8 above
Annex 1	This annex on antiestrogenic activity testing should either be more specific or excluded until such time as an appropriately validated protocol that measures such activity can be delineated.